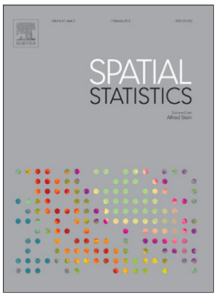


Shape-restricted estimation and spatial clustering of COVID-19 infection rate curves

James Matuk, Xiaohan Guo

PII: S2211-6753(21)00056-7
DOI: https://doi.org/10.1016/j.spasta.2021.100546
Reference: SPASTA 100546

To appear in: *Spatial Statistics*

Received date : 27 February 2021
Revised date : 28 September 2021
Accepted date : 3 October 2021

Shape-Restricted Estimation and Spatial Clustering of COVID-19 Infection Rate Curves

James Matuk[a,*], Xiaohan Guo[b]

[a]Duke University, 214 Old Chemistry, Durham, NC 27708, USA
[b]The Ohio State University, 281 W Lane Ave, Columbus, OH 43210, USA

ARTICLE INFO

Keywords:
COVID-19
Spatio-Temporal modeling
Functional Data Analysis
Bayesian Inference

ABSTRACT

The study of regional COVID-19 daily reported cases is used to understand pattern of spread and disease progression over time. These data are challenging to model due to noise that is present, which arises from failures in reporting, false positive tests, etc., and the spatial dependence between regions. In this work, we extend a recently developed Bayesian modeling framework for inference of functional data to jointly estimate and cluster daily reported cases data from US states, while accounting for spatial dependence between US states. Shape-restriction allows us to directly infer the number of extrema of a smooth infection rate curve that underlies noisy data. Other parameters in the model account for the relative timing of extrema, and the magnitude and severity of infection rates. We incorporate mobility behavior of each US state's population into an informative prior model to account for the spatial dependence between US states. Our model corroborates past work that shows that different US states have indeed experienced COVID-19 differently, but that there are regional patterns within the US. The modeling results can be used to assess severity of infection in individual US states and trends of neighboring US states to aid pandemic planning. Retrospectively, this model can be used to see which factors (governmental, behavioral, etc.) are associated with the varying shapes of infection rate curves, which is left as future work.

1. Introduction

In late 2019, the outbreak of Coronavirus Disease - 2019 (COVID-19) was reported in Wuhan, China, and the disease quickly spread worldwide in early 2020 (Wu, Chen and Chan, 2020). The United States (US), which is the focus of our work, has been one of the most affected countries in the world. Within months of the first reported cases in the US, more infections were reported than in any other country, and the disease has remained a major problem for the health and safety of its citizens (Omer, Malani and del Rio, 2020). Unmitigated growth of infections could potentially over-burden hospital resources, and poses a severe threat to individuals in populations associated with increased risk of serious illness and mortality (Wolff, Nee, Hickey and Marschollek, 2020). These facts have brought epidemiological modeling of COVID-19 to the foreground in hopes that understanding disease progression, transmission, and pervasiveness will inform public health interventions (Thompson, 2020; Pei, Kandula and Shaman, 2020).

Bertozzi, Franco, Mohler, Short and Sledge (2020) state that the novelty and dynamic nature of COVID-19 are the primary challenges in effective modeling, and offer a survey of modeling approaches. Many current models rely on the compartmental Susceptible-Infected-Recovered (SIR) models and its variants (Cooper, Mondal and Antonopoulos, 2020; Arenas, Cota, Gómez-Gardeñes, Gómez, Granell, Matamalas, Soriano and Steinegger, 2020; Giordano, Blanchini, Bruno, Colaneri, Filippo, Matteo and Colaneri, 2020; Sharma, Volpert and Banerjee, 2020; Xue, Jing, Miller, Sun, Li, Estrada-Franco, Hyman and Zhu, 2020). SIR models directly model the mechanism of infection and spread, and consequently, are able to explicitly infer important quantities, such as reproductive numbers, which quantify the contagiousness of a disease. As the pandemic progresses, there is a need to understand dynamic changes in infection rates as well as differences in geographic regions. Traditional SIR models do not specialize in capturing the flexible dynamic trend of pandemic curves and can ignore important spatial information, resulting in limited applicability in spatio-temporal modelling and potential biased inference due to model misspecification, especially for a novel disease.

*Corresponding author
james.matuk@duke.edu (J. Matuk); guo.1280@osu.edu (X. Guo)
https://jamesmatuk.weebly.com/ (J. Matuk)
ORCID(s): 0000-0002-1720-9092 (J. Matuk)

In this work, we focus on modelling spatial patterns and temporal trends of infection rate curves. In contrast with SIR models, Functional Data Analysis (FDA) offers more flexible approaches for understanding trends (Boschi, Iorio, Testa, Cremona and Chiaromonte, 2020; Srivastava and Chowell, 2021). The flexibility of FDA enables the incorporation of spatial information into modeling, and allows for spatial information to support various statistical analysis tasks (Pan, Shen and Hu, 2020). Consequently, FDA approaches can be used to account for important aspects of modeling the disease, such as spatial heterogeneity (Thomas, Huang, Yin, Luo, Almquist, Hipp and Butts, 2020) and spatial dependence (Guliyev, 2020). COVID-19 has manifested differently in regions within the same country which has influenced awareness and response to the disease. The dependence among transmission dynamics of states is related to inter-state transportation and similarity in socioeconomic factors, climate, and public health policy, which depends on the spatial distribution of states (Badr, Du, Marshall, Dong, Squire and Gardner, 2020; Cintia, Pappalardo, Rinzivillo, Fadda, Boschi, Giannotti, Chiaromonte, Bonato, Fabbri, Penone et al., 2020). We propose a FDA model that accounts for both spatial heterogeneity and dependence.

Approaches from FDA use smooth infection rate curves based on daily reported cases to study regional infection rate dynamics through cluster analysis. With the exception of Srivastava and Chowell (2021), these works tend to overlook the distinct forms of variability in functional data: amplitude, which quantifies features of functions, e.g. magnitude and number of extrema, and phase, which quantifies the relative timing of the amplitude features. Within the context of modeling US states' infection rate curves, amplitude and phase components provide complementary information about how each US state has experienced the pandemic. The study of amplitude enables an interpretation of the magnitude and pattern of spread of the infection in each US state. On the other hand, the study of phase quantifies differences in when US states experienced ebbs and flows of infections.

Matuk, Bharath, Chkrebtii and Kurtek (2021) recently developed a Bayesian modeling framework for estimation of different types of functional data in the presence of phase variation. In this work, we extend their framework to model local extrema information of infection rate curves from daily reported COVID-19 cases while accounting for inter-state spatial dependence. Our primary contributions are summarized as follows:

i. Our use of the Bayesian paradigm allows us to jointly model three types of variation in COVID-19 infection rate data: amplitude, phase, and spatial variation. The latent amplitude and phase components are modelled as functional parameters, through a formulation rooted in the Elastic Functional Data Analysis (EFDA) framework (Srivastava, Wu, Kurtek, Klassen and Marron, 2011). Spatial dependence is introduced through judicious choice of prior distributions for model parameters. Moreover, the Bayesian framework allows for structured uncertainty quantification of all of these aspects of the model.

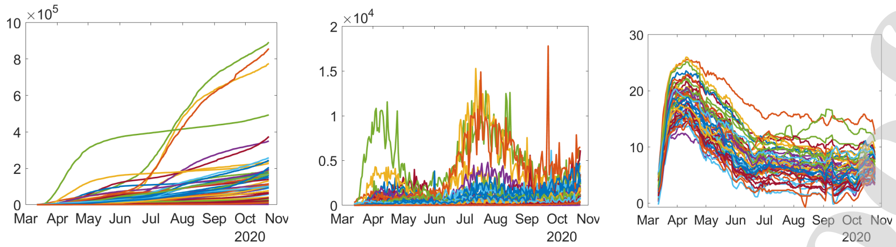
ii. The proposed spatial prior not only includes information about the neighborhood structure among US states, but also important covariate information in terms of community mobility data. We implement amplitude-phase separation in defining a variogram to estimate a spatial correlation measure.

iii. We take a shape-restricted perspective for inference which enables the analysis of amplitude and phase variability in the presence of noisy data. Further, the model allows us to infer the number and pattern of extrema of infection rate curves. As we show in Section 4, extrema are useful in understanding the pattern of infections within US states, and heterogeneity of infections between US states. We use extrema to inform spatial clustering, which unveils regional patterns of COVID-19 infection rates throughout the US. The parameters for phase quantify differences in timing of extrema of infection rate curves and also contribute to our spatial clustering results.

The remainder of this paper is organized as follows. In Section 2, we discuss the data that we use to inform our model. In Section 3, we briefly discuss the EFDA framework which serves as the theoretical foundation of our model, and formulate our model. In Section 4, we discuss our modeling results. In Section 5, we summarize our contributions and discuss future work.

2. Data Description

We model daily reported cases data calculated from cumulative reported cases data compiled by The COVID Tracking Project. The COVID Tracking Project is a volunteer organization that is dedicated to manually compiling data from US states' COVID-19 dashboards and making them publicly available (Covid Volunteer Team, 2020). Figure 1 panel (a) shows the cumulative reported cases data for the 50 US states and Washington DC, which we simply refer to as 'US states' in our study, and panel (b) shows daily reported cases data for these US states, which are used to



(a) (b) (c)

Figure 1: (a) Number of cumulative reported cases in each US state and Washington DC. (b) Number of daily reported cases in each US state and Washington DC. (c) Mobility data, smoothed using a 7-day moving average, that represent percentage change from baseline for typical time spent in a residential area of each US states' population.

inform our model. As is shown in Section 4, the infection rate curves that underlie the noisy data in panel (b) have different patterns of extrema in terms of number and location, with regional similarities in the shapes of these curves.

We incorporate spatial dependence into our model through the use of informative prior probability modeling that accounts for correlation between neighboring US states. In our model, the spatial correlations between US states is determined by geographic distance between states and also the similarity in US states' mobility behavior. This information is important to be included in our model, since early studies of the COVID-19 pandemic in the US and other countries (Badr et al., 2020; Kraemer, Yang, Gutierrez, Wu, Klein, Pigott, Du Plessis, Faria, Li, Hanage et al., 2020; Yang, Sha, Liu, Li, Lan, Guan, Hu, Li, Zhang, Thompson et al., 2020) suggested that mobility is an essential factor associated with infection rates. To quantify mobility, we use the community mobility data collected by Google (Google LLC, 2020). The mobility reports track daily percentage changes in the duration of time spent in residential areas relative to a pre-pandemic baseline for the population in each US state. We display smoothed versions of these data in Figure 1 panel (c). These mobility data take the form of functional observations that capture similarities in how US states' populations responded to the pandemic and the ensuing public health policies. We find them useful for determining the spatial dependence in our model as discussed in Section 3.3.

Throughout this work, all data was recorded starting from March 12th, 2020, the day after the pandemic was declared by the World Health Organization (Ghebreyesus, 11 March 2020), through October 25th, 2020. As is often done in practice (Ramsay and Silverman, 2005; Srivastava and Klassen, 2016), the different days throughout this span are represented as time points on the unit interval, $t \in [0, 1]$. Further discussion of these data and pre-processing are discussed in Section 1 of the Supplemental Material.

3. Methods

In this section, we briefly summarize the EFDA framework, which serves as the theoretical background of our model, followed by our model specification for COVID-19 infection rate curves.

3.1. Elastic Functional Data Analysis Background

As briefly mentioned in Section 1, one of the primary challenges in working with functional data is the presence of amplitude and phase variability (Marron, Ramsay, Sangalli and Srivastava, 2015). Within the context of COVID-19 infection rate curves, there are likely many factors associated with the presence of these variabilities, including the date when cases were first reported, the latency of reporting due to limited testing capacity in early stages of the pandemic, governmental response to perceived surges in infections, etc. While there are many approaches to addressing the issues associated with these types of variabilities (Ramsay and Silverman, 2005; Srivastava and Klassen, 2016), we choose to work within the EFDA framework (Srivastava et al., 2011), which accounts for amplitude and phase variability through a registration procedure with desirable properties.

The registration procedure decomposes a functional dataset, $f_1, \dots, f_n$, into a set of amplitude functions, $\tilde{f}_1, \dots, \tilde{f}_n$, and phase functions, $\gamma_1, \dots, \gamma_n$, such that their composition is equal to the original functional dataset , $f_i = \tilde{f}_i \circ \gamma_i$ $i = 1, \dots, n$. The phase functions are elements of the group of diffeomorphisms on the unit interval, $\Gamma = \{\gamma : [0, 1] \to [0, 1] \mid \gamma(0) = 0, \gamma(1) = 1, \dot{\gamma} > 0\}$, where $\dot{f}$ represents the time derivative of a function f. For identifiability, the sample Karcher mean of the phase functions is forced to be the identity warping, $\gamma_{id}(t) = t$. The optimality criterion used to determine the registration is based in the Fisher-Rao Riemannian metric on the space of absolutely continuous functions. Square root velocity function (SRVF) representation is a crucial ingredient of the EFDA framework, since it enables simple computation of the Fisher-Rao Riemannian metric. The SRVF of a function, f, is defined as, $q = Q(f) := \dot{f}/\sqrt{|\dot{f}|}$. This representation is invertible, if in addition to the SRVF, q, one also records the starting point on the function, f(0), $Q^{-1}(q, f(0))(t) := f(0) + \int_0^t q(s)|q(s)|ds$. Our model, presented in the next subsection, relies on SRVF representation to decompose amplitude and phase of infection rate curves through model parameters.

3.2. Shape-restricted Amplitude Model

Let $\mathbf{y}_i$ denote the daily reported cases, as discussed in Section 2, and let $\mathbf{t}_i$ represent days where non-missing values were recorded for the i^{th} US state, $i = 1, \ldots, 51$. Throughout this section, we use the notation $f(\mathbf{t}_i)$ to denote a vector of function evaluations $(f(t_{i,1}), \ldots, f(t_{i,m_i}))^{\top}$, where m_i is the length of $\mathbf{t}_i$. We assume the following observation model,

$$\mathbf{y}_i = (Q^{-1}(q_i^{(H_i,M_i)}, T_i) \circ \gamma_i)(\mathbf{t}_i) + \epsilon_i(\mathbf{t}_i), \quad i = 1, \ldots, 51. \tag{1}$$

This specifies the daily reported cases data for the i^{th} US state as perturbations of a smooth infection rate curve, $(Q^{-1}(q_i^{(H_i,M_i)}, T_i) \circ \gamma_i)$, whose components $q_i^{(H_i,M_i)}$, γ_i, T_i denote amplitude, phase, and translation, respectively. The amplitude component $q_i^{(H_i,M_i)}$ is indexed by the parameters H_i and M_i that determine the number and ordering of extrema of the function $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$.

We represent the amplitude component $q_i^{(H_i,M_i)}$ via an expansion of shape-restricted basis functions (Wheeler, Dunson and Herring, 2017), $\{W_b^{(H_i,M_i)}\}_{b=1}^{B}$,

$$q_i^{(H_i,M_i)} = \sum_{b=1}^{B} \beta_{i,b} W_b^{(H_i,M_i)}. \tag{2}$$

The basis elements are defined as $W_b^{(H_i,M_i)}(t) = M_i(\prod_{h=1}^{H_i}(t - \alpha_h^{(H_i)}))U_b(t)$, $b = 1, \ldots, B$, where $U_b(t)$ are B-spline basis functions, and the basis coefficients are assumed to follow a weakly-informative exponential prior, $\beta_{i,b} \overset{\text{iid}}{\sim} \exp(\lambda_\beta)$. In this work, we select a value of $B = 10$ based on exploratory data analysis to reflect a flexible, yet parsimonious basis. As discussed in Matuk et al. (2021), the shape-restricted basis is robust to the choice of B for fixed values of (H_i, M_i). This basis representation forces the amplitude component of the model, $q_i^{(H_i,M_i)}$, to be zero at the values of $\alpha^{(H_i)}$, which is a vector of H_i evenly spaced points throughout the domain. The zeros of $q_i^{(H_i,M_i)}$ correspond to extrema for the function $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$. The parameter M_i determines the ordering of extrema. For example, when $H_i = 2$, $M_i = 1$ enforces $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$ to have a local maximum followed by a local minimum, while $M_i = -1$ enforces $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$ to have a local minimum followed by a local maximum.

Since infection curves for states exhibit variations in numbers of peaks and valleys, we extend the model by Matuk et al. (2021), where H_i and M_i are fixed a-priori, to allow for H_i and M_i to take on a range of values. We specify a novel prior for these parameters as,

$$p(H_i = h | \{H_j\}_{j \neq i}) \propto (1 - \nu) + \nu \frac{\sum_{j \neq i:\ H_j = h} \omega_{i,j}}{K}, \quad h = 0,\ 1, \ldots, H_{max}, \tag{3}$$

$$M_i = \begin{cases} 1 & H_i \text{ is even} \\ -1 & H_i \text{ is odd .} \end{cases} \tag{4}$$

The number of extrema for the infection rate curve of the i^{th} US state, H_i, is able to vary between 0 to H_{max}, and its probable values depend on nearby US states. The value of $H_{max} = 5$ enforces the assumption that no state has experienced more than 3 distinct waves of infections throughout the time that the data has been collected, which was chosen based on exploratory data analysis of the observations. In general, specification of the parameter H_{max} too small can result in oversmoothing features of the data, and H_{max} too large can result in mistaking noise as extrema. M_i is specified so that $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$ is increasing in the beginning of the domain.

The dependence between H_i and H_j is measured by a spatial weight, $\omega_{i,j}$, where the weights are truncated so that only bordering US states have a positive weight. These weights depend both on geographic distance and similarity in the changes of US states populations' mobility behavior during the pandemic. The methodology for determining the weights is discussed in Section 3.3. The constant $K = 8$ corresponds to the largest number of bordering US states for any US state, and normalizes the spatial component of the prior to a value between 0 and 1. The regularization parameter, $\nu \in [0, 1]$, determines the strength of regularization. When $\nu = 0$, the number of extrema in the smooth daily infection rate curve for each US state follows a discrete uniform prior that does not depend on neighboring US states, and as ν approaches 1, the prior on H_i, the number of extrema in the smooth daily infection rate curve for state i, will favor the numbers of extrema that are similar to those of its neighbors. In this work, we select a moderate value for $\nu = 0.99$ using sensitivity analysis to strike a balance between incorporating spatial information into the model and over-regularization. Section 2 in the Supplemental Material discusses the sensitivity analysis and corresponding selection of this parameter.

The phase component of the i^{th} US state, γ_i, shifts the extrema of the function $Q^{-1}(q_i^{(H_i,M_i)}, T_i)$ to fit the the daily reported case data. The prior is defined through the finite difference of a discretized phase function,

$$p(\gamma_i(\mathbf{t}_\gamma)) := (\gamma_i(t_{\gamma,1}), \dots, \gamma_i(t_{\gamma,k}) - \gamma_i(t_{\gamma,k-1}), \dots, \gamma_i(t_{\gamma,m_\gamma}) - \gamma_i(t_{\gamma,m_\gamma-1}))^\top \sim \text{Dirichlet}(\theta_\gamma p(\gamma_{id}(\mathbf{t}_\gamma))), \tag{5}$$

proposed by Bharath and Kurtek (2020), where $\mathbf{t}_\gamma$ is a grid of evenly space points along the domain and θ_γ is a concentration hyperparameter that we choose to correspond to a diffuse prior. We assume an independent and identically distributed $\epsilon(\mathbf{t}_i) \sim N_{m_i}(\mathbf{0}_{m_i}, \sigma_i^2 I_{m_i})$, which elicits a normal likelihood, and we model the translation and noise variance for each US state with weakly-informative normal and inverse-gamma conjugate priors, respectively.

Posterior inference is based on Markov chain Monte Carlo (MCMC) samples from an adaptive parallel tempering algorithm (Strait, Chkrebtii and Kurtek, 2019), where the temperature scheme (Geyer, 1991) and the proposal parameters of the Metropolis-within-Gibbs algorithm (Roberts and Rosenthal, 2009) are automatically tuned for efficiency. We summarize our hierarchical model and discuss MCMC implementation in Section 3 of the Supplemental Material.

3.3. Spatial Weights Informed by Mobility Data

Motivated by the association between infection rate and community mobility, spatial correlation between US states' mobility data is used to inform the dependence between the number of extrema H_i in infection rate curves. Modeling spatial dependence of mobility data for constructing a prior distribution of H_i is challenging, since the state-level weekly community mobility records are functional data that exhibit phase variation. The presence of phase variation can blur the inter-state dependence structure, and can cause biased inference when ignored. The shape parameter H_i is only related to the fluctuation of infection rate curves but invariant to phase variation, so we desire that the estimated inter-state correlation of mobility data is also invariant to potential phase variation. To satisfy this requirement, the amplitude trace-variogram approach proposed by Guo, Kurtek and Bharath (2020) is implemented to measure the spatial variation between the amplitudes of mobility curves.

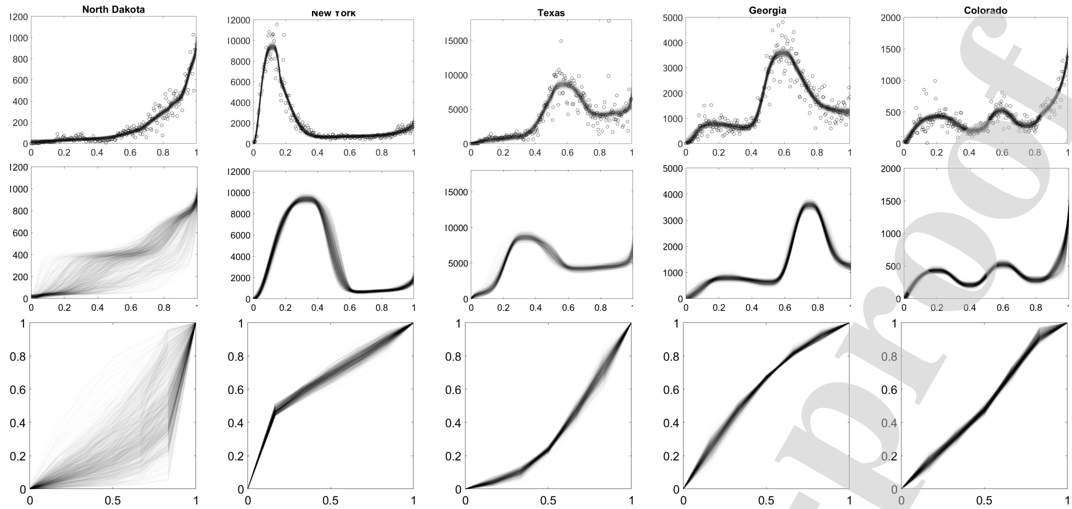
The amplitude trace-variogram is defined based on amplitude-phase decomposition of a trace-variogram, which is a popular spatial variation measure for functional data (Giraldo, Delicado and Mateu, 2011). We model a random field that takes on functional values as a coupling of a latent amplitude random field and phase random field. We estimate the amplitude trace-variogram through incorporating function registration into variogram estimation. This procedure provides spatial weights that capture correlation between magnitudes of mobility records without interference of phase variation potentially caused by various factors, such as lock-down policies among states.

In the functional space $\mathcal{G} = \{g : [0, 1] \mapsto R \mid g \textit{ is absolutely continuous}\}$, consider the functional random field $\{g_{\mathbf{s}} \mid \mathbf{s} \in \mathcal{D}\} \subset \mathcal{G}$ on a spatial coordinate domain $\mathcal{D} \subset R^2$, where at each $\mathbf{s} \in \mathcal{D}$, $g_{\mathbf{s}}$ is a random function in $\mathcal{G}$. Suppose the mobility data $g_{\mathbf{s}_i}$ is one realization of the functional random field in the US state located at $\mathbf{s}_i$, for $i = 1, \dots, 51$. Using the SRVF representation, the model for the mobility data is,

$$g_{\mathbf{s}_i} = \left[Q^{-1}(\mu + e_{\mathbf{s}_i})\right] \circ \xi_{\mathbf{s}_i}^{-1}(t),\ t \in [0, 1],\ i = 1, \dots, 51, \tag{6}$$

where $\mu(t)$ is the unobserved deterministic mean amplitude, constant in space; $e_{\mathbf{s}_i}(t)$ is the random process with mean 0 and covariance function $\mathcal{K}(t_1, t_2)$ s.t. $\int_0^1 \mathcal{K}(t, t)dt < \infty$; $\xi_{\mathbf{s}_i} \in \Gamma$ is the unobserved random phase component with $E(\xi_{\mathbf{s}_i}) = \xi \in \Gamma$ for all $\mathbf{s}_i \in \mathcal{D}$. The inter-state dependence structure of interest is contained in amplitude error $e_{\mathbf{s}_i}(t)$, which is not directly estimable due to the existence of unknown warping $\xi_{\mathbf{s}_i}(t)$. Equivalently, the amplitude model is,

$$Q(g_{\mathbf{s}_i} \circ \xi_{\mathbf{s}_i})(t) = \mu(t) + e_{\mathbf{s}_i}(t),\ t \in [0, 1],\ i = 1, \dots, 51. \tag{7}$$



(a) (b) (c) (d) (e)

Figure 2: (a)-(e) Posterior draws (transparent lines) representing phase (bottom) amplitude (middle) and their composition (top) with the observations (dots) superimposed for 5 US states in different clusters.

The amplitude random field is defined as $\{Q(g_{\mathbf{s}} \circ \xi_{\mathbf{s}}) \mid \mathbf{s} \in \mathcal{D}\}$. Under the second-order stationary and isotropic assumptions on $\{Q(g_{\mathbf{s}} \circ \xi_{\mathbf{s}}) \mid \mathbf{s} \in \mathcal{D}\}$, the amplitude trace-varioram is defined as a function of spatial distance,

$$V(\|\mathbf{s} - \mathbf{s}'\|) = \frac{1}{2} E\left(\|Q(g_{\mathbf{s}} \circ \xi_{\mathbf{s}}) - Q(g_{\mathbf{s}'} \circ \xi_{\mathbf{s}'})\|^2\right), \tag{8}$$

where $\|\cdot\|$ is the L^2-norm. The corresponding empirical amplitude trace-variogram (Guo et al., 2020), based on an amplitude distance defined by optimising over Γ (Srivastava and Klassen, 2016), is defined as,

$$\widetilde{V}(r) = \frac{1}{2|N(r)|} \sum_{i,j \in N(r)} \inf_{\xi \in \Gamma} \|Q(g_{\mathbf{s}_i} \circ \xi) - Q(g_{\mathbf{s}_j})\|, \ r \geq 0,$$

where $N(r) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in (r - \delta, r + \delta)\}$ for a small $\delta > 0$. To guarantee the estimated variogram is conditionally negative definite, we further fit a parametric Matérn variogram to the empirical variogram $\tilde{V}(r)$, resulting in a parametric estimate $\widehat{V}(r)$. In model fitting, the smoothness parameter is fixed at 0.5, while the range ϕ and sill τ are estimated through least squares. Finally, we specify the spatial weight as,

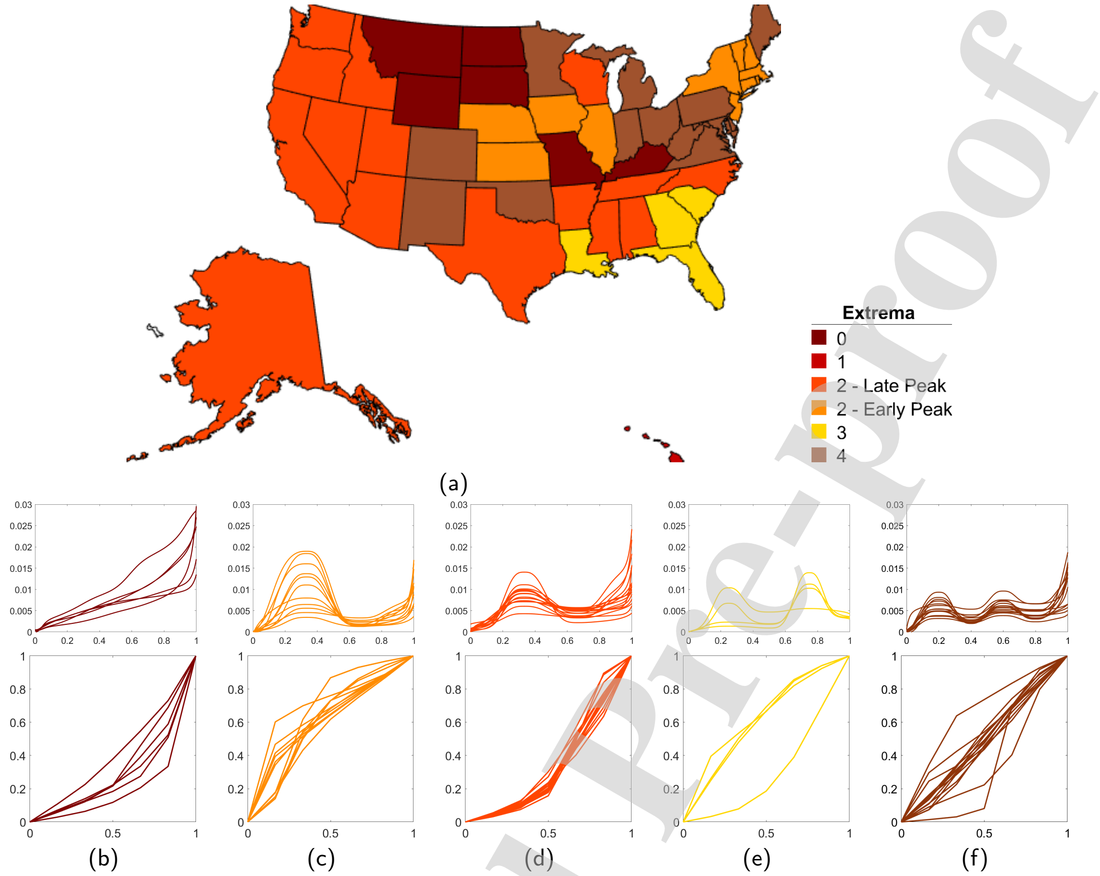
$$\omega_{ij} = 1 - \widehat{V}(\|\mathbf{s}_i - \mathbf{s}_j\|)/\hat{\tau}, \ i, j = 1, \ldots, 51, \tag{9}$$

where $\hat{\tau}$ is the estimated scale parameter of the Matérn family. We note that ω_{ij} is in $[0, 1]$ and is 0 when $\|\mathbf{s}_i - \mathbf{s}_j\| \geq \hat{\phi}$.

4. Results

In this section, we present our modeling results for infection rate curves based on US states' daily reported cases data. Throughout this section, $\widehat{H}_i$ is used to denote the estimated marginal posterior mode of the number of extrema for the i^{th} US state, and functional posterior samples and summaries are shown given this value. Clusters are determined by the estimated number of extrema. Analysis of estimated posterior phase functions, used to quantify the relative timing of extrema, are useful for identifying sub-clusters.

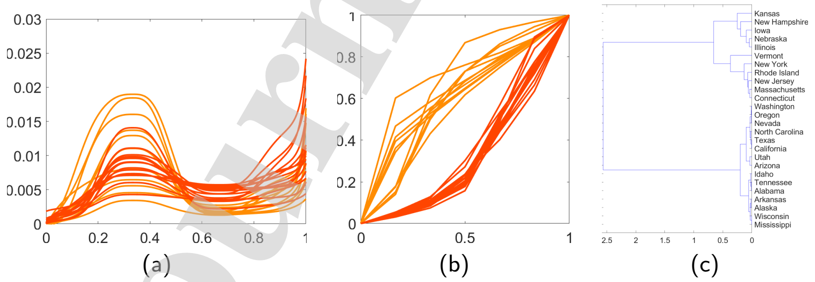
Figure 2 shows results for some selected US states that are representative of the patterns seen in the estimated infection rate curves as determined by our model. In each of the panels, the dots represent the daily reported cases data for each of the US states, and the gray lines represent marginal posterior draws of functional parameters from our model. The top row shows the daily reported cases data superimposed on posterior draws of the infection rate curve



(a)

(b) (c) (d) (e) (f)

Figure 3: (a) A map of the US colored by cluster membership. (b)-(f) Cluster summaries representing posterior mean amplitude for (top) and phase (bottom) estimated from MCMC draws. Note: the amplitude functions are rescaled by the total number of reported cases in the corresponding US state to aid visualization.



(a) (b) (c)

Figure 4: (a) Posterior mean amplitude functions with two extrema colored by phase subcluster membership. (b) Posterior mean phase functions corresponding to the amplitude functions in panel (a) colored by subcluster membership. (c) Dendrogram representing hierarchical clustering results based on the posterior mean phase functions. Note: the amplitude functions are rescaled by the total number of reported cases in the corresponding US state to aid visualization.

parameter given the estimated marginal posterior mode of the number of extrema present in the infection rate curve, $\hat{H}_i$. Regardless of the pattern that is present for each of the US states, the model appears to reasonably estimate a smooth infection rate curve. These results represent the heterogeneity of shapes that are present in estimated infection rate curves. The infection rate for North Dakota has only steadily increased throughout the time of the study; US states like New York and Texas have experienced one prominent peak followed by increasing cases; Georgia and Colorado have both experienced two distinct peaks in their infection rate curves, but Colorado's current infection rate is increasing, while Georgia's is currently decreasing. The middle and bottom rows of the figure display the amplitude and phase decomposition of the infection rate curves. The middle row shows marginal posterior amplitude functions, which are aligned for US states with the same number of extrema, $\hat{H}_i$. The bottom row of the figure displays posterior draws for the phase component, which is able to quantify the relative timing of extrema. This is apparent when comparing the results in panels (b) and (c). The model estimates that both New York and Texas have two extrema, a peak followed by a valley. However, the marginal posterior phase functions for New York are generally higher than that of Texas, indicating that the infection peak of New York was experienced much sooner. Refer to Section 2 of the supplemental material to view the effect of the regularization parameter, ν, on estimation results. The parameter ensures that the pattern of extrema for estimated functions are influenced by their neighbors, which in some cases can prevent overfitting.

Figure 3 shows a map of US states colored by cluster memberships determined by the estimated number and timing of extrema in panel (a) along with estimated marginal posterior mean amplitude and phase functions of US states within each cluster in panels (b)-(f). These results indicate that there is strong regional behavior for the shape of infection rate curves in the US. The most prominent regional clusters are the maroon US states in the Northwest, the brown US states in the Atlantic and Midwest region, the yellow US states in the south, and US states in different shades of orange in the Northeast and West. The US states in different shades of orange have the same number of extrema, however, the times at which the extrema occur separates the groups, characterized by marginal posterior phase mean functions. The group colored in light orange has peaks concentrated around the end of March and early-April and the group colored in dark orange has peaks concentrated in mid-June. This subclustering based on phase is investigated further in Figure 4. The figure again shows the marginal posterior mean amplitude and phase functions for the US states in the orange cluster to highlight the differences in the phase functions. A dendrogram visualizing an agglomerative hierarchical clustering with Ward linkage (Köhn and Hubert, 2015), based on all of the posterior mean phase functions for US states with two extrema, confirms that there are two prominent groups. Of all of the clusters based solely on the number of extrema, we have determined that this is the most appropriate to subcluster based on phase.

5. Discussion

In this paper, we have presented a model to infer infection rate curves in the US from noisy daily reported cases data. Importantly, this approach provides a flexible model for the spatially correlated, noisy data that directly infers the number and timing of extrema of underlying smooth functions, which are crucial in visualizing and understanding the regional patterns of COVID-19 within the US.

In comparison with Srivastava and Chowell (2021), which also uses EFDA for clustering of infection rate curves, our approach has some key methodological differences. We are able to incorporate spatial information for inference through prior modeling of parameters. We base clustering on the quantity and relative timing of extrema, which aids the interpretation of clusters compared to unsupervised learning of latent groups based on pairwise distances between functions. Our model performs simultaneous estimation, registration, and clustering of observations in contrast with performing these steps sequentially, which fails to propagate uncertainty from the estimation step to the clustering step.

The proposed model can be used to assess current severity of infection in individual US states and trends of neighboring US states to help with pandemic planning. In future work, we believe this model could be adapted within a functional regression framework to determine which factors (governmental, behavioral, etc.) are associated with the varying shapes of infection rate curves. This model could be modified for the analysis of county-level data, where the high number of zero reported cases in rural counties would need to be explicitly accounted for through the use of a zero-inflated likelihood model.

6. Acknowledgements

The authors would like to thank Dr. Karthik Bharath (University of Nottingham), Dr. Oksana Chkrebtii (The Ohio State University), and Dr. Sebastian Kurtek (The Ohio State University) for their helpful conversations. This research was partially supported by grants NSF CCF-1740761 and NSF CCF-1839252.

References

Arenas, A., Cota, W., Gómez-Gardeñes, J., Gómez, S., Granell, C., Matamalas, J.T., Soriano, D., Steinegger, B., 2020. A mathematical model for the spatiotemporal epidemic spreading of covid-19. medRxiv URL: https://www.medrxiv.org/content/early/2020/03/23/2020.03.21.20040022, doi:10.1101/2020.03.21.20040022, arXiv:https://www.medrxiv.org/content/early/2020/03/23/2020.03.21.20040022.full.pdf.

Badr, H.S., Du, H., Marshall, M., Dong, E., Squire, M.M., Gardner, L.M., 2020. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. The Lancet Infectious Diseases 20, 1247–1254.

Bertozzi, A.L., Franco, E., Mohler, G., Short, M.B., Sledge, D., 2020. The challenges of modeling and forecasting the spread of covid-19. Proceedings of the National Academy of Sciences 117, 16732–16738. URL: https://www.pnas.org/content/117/29/16732, doi:10.1073/pnas.2006520117, arXiv:https://www.pnas.org/content/117/29/16732.full.pdf.

Bharath, K., Kurtek, S., 2020. Distribution on warp maps for alignment of open and closed curves. Journal of the American Statistical Association 115, 1378–1392. URL: https://doi.org/10.1080/01621459.2019.1632066, doi:10.1080/01621459.2019.1632066, arXiv:https://doi.org/10.1080/01621459.2019.1632066.

Boschi, T., Iorio, J.D., Testa, L., Cremona, M.A., Chiaromonte, F., 2020. The shapes of an epidemic: using functional data analysis to characterize covid-19 in italy. arXiv:2008.04700.

Cintia, P., Pappalardo, L., Rinzivillo, S., Fadda, D., Boschi, T., Giannotti, F., Chiaromonte, F., Bonato, P., Fabbri, F., Penone, F., et al., 2020. The relationship between human mobility and viral transmissibility during the covid-19 epidemics in italy. arXiv preprint arXiv:2006.03141 .

Cooper, I., Mondal, A., Antonopoulos, C., 2020. A sir model assumption for the spread of covid-19 in different communities. Chaos, Solitons & Fractals 139, 110057. doi:10.1016/j.chaos.2020.110057.

Covid Volunteer Team, 2020. Covid tracking project. URL: https://covidtracking.com/.

Geyer, C., 1991. Markov chain Monte Carlo maximum likelihood, in: Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface, 156, American Statistical Association.

Ghebreyesus, T.A., 11 March 2020. Who director-general's opening remarks at the media briefing on covid-19.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Filippo, A., Matteo, A., Colaneri, M., 2020. Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. Nature Medicine 26, 1–6. doi:10.1038/s41591-020-0883-7.

Giraldo, R., Delicado, P., Mateu, J., 2011. Ordinary kriging for function-valued spatial data. Environmental and Ecological Statistics 18, 411–426.

Google LLC, 2020. Google covid-19 community mobility reports. URL: https://www.google.com/covid19/mobility/.

Guliyev, H., 2020. Determining the spatial effects of covid-19 using the spatial panel data model. Spatial statistics 38, 100443.

Guo, X., Kurtek, S., Bharath, K., 2020. Variograms for spatial functional data with phase variation. arXiv:2010.09578.

Kraemer, M.U., Yang, C.H., Gutierrez, B., Wu, C.H., Klein, B., Pigott, D.M., Du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., et al., 2020. The effect of human mobility and control measures on the covid-19 epidemic in china. Science 368, 493–497.

Köhn, H.F., Hubert, L.J., 2015. Hierarchical Cluster Analysis. John Wiley & Sons, Ltd. pp. 1–13. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat02449.pub2, doi:https://doi.org/10.1002/9781118445112.stat02449.pub2, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat02449.pub2.

Marron, J.S., Ramsay, J.O., Sangalli, L.M., Srivastava, A., 2015. Functional data analysis of amplitude and phase variation. Statistical Science 30, 468–484.

Matuk, J., Bharath, K., Chkrebtii, O., Kurtek, S., 2021. Bayesian framework for simultaneous registration and estimation of noisy, sparse and fragmented functional data. Journal of the American Statistical Association URL: https://doi.org/10.1080/01621459.2021.1893179, doi:10.1080/01621459.2021.1893179, arXiv:https://doi.org/10.1080/01621459.2021.1893179.

Omer, S.B., Malani, P., del Rio, C., 2020. The covid-19 pandemic in the us. Journal of the American Medical Association 323. URL: https://jamanetwork.com/journals/jama/fullarticle/2764366, doi:10.1001/jama.2020.5788.

Pan, T., Shen, W., Hu, G., 2020. Spatial homogeneity learning for spatially correlated functional data with application to covid-19 growth rate curves. arXiv:2008.09227.

Pei, S., Kandula, S., Shaman, J., 2020. Differential effects of intervention timing on covid-19 spread in the united states. Science Advances 6. URL: https://advances.sciencemag.org/content/6/49/eabd6370, doi:10.1126/sciadv.abd6370, arXiv:https://advances.sciencemag.org/content/6/49/eabd6370.full.pdf.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer.

Roberts, G.O., Rosenthal, J.S., 2009. Examples of adaptive mcmc. Journal of Computational and Graphical Statistics 18, 349–367. URL: https://doi.org/10.1198/jcgs.2009.06134, doi:10.1198/jcgs.2009.06134, arXiv:https://doi.org/10.1198/jcgs.2009.06134.

Sharma, S., Volpert, V., Banerjee, M., 2020. Extended seiqr type model for covid-19 epidemic and data analysis. Mathematical biosciences and engineering: MBE 17. doi:10.3934/mbe.2020386.

Srivastava, A., Chowell, G., 2021. Title: Modeling study: Characterizing the spatial heterogeneity of the covid-19 pandemic through shape analysis of epidemic curves. doi:10.21203/rs.3.rs-223226/v1.

Srivastava, A., Klassen, E., 2016. Functional and Shape Data Analysis. Springer.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., Marron, J.S., 2011. Registration of functional data using fisher-rao metric. arXiv:1103.3817.

Strait, J., Chkrebtii, O., Kurtek, S., 2019. Parallel tempering strategies for model-based landmark detection on shapes. Communications in Statistics - Simulation and Computation 0, 1–21. URL: https://doi.org/10.1080/03610918.2019.1670843, doi:10.1080/03610918.2019.1670843, arXiv:https://doi.org/10.1080/03610918.2019.1670843.

Thomas, L.J., Huang, P., Yin, F., Luo, X.I., Almquist, Z.W., Hipp, J.R., Butts, C.T., 2020. Spatial heterogeneity can lead to substantial local variations in covid-19 timing and severity. Proceedings of the National Academy of Sciences of the United States of America 117. doi:10.1073/pnas.2011656117.

Thompson, R., 2020. Epidemiological models are important tools for guiding covid-19 interventions. BMC Medicine 18, 152. doi:10.1186/s12916-020-01628-4.

Wheeler, M.W., Dunson, D.B., Herring, A.H., 2017. Bayesian local extremum splines. Biometrika 104, 939–952. URL: https://doi.org/10.1093/biomet/asx039, doi:10.1093/biomet/asx039, arXiv:https://academic.oup.com/biomet/article-pdf/104/4/939/21813603/asx039.pdf.

Wolff, D., Nee, S., Hickey, N., Marschollek, M., 2020. Risk factors for covid-19 severity and fatality: a structured literature review. Infection doi:10.1007/s15010-020-01509-1.

Wu, Y.C., Chen, C.S., Chan, Y.J., 2020. The outbreak of covid-19: An overview. Journal of the Chinese Medical Association 83, 217–220. doi:10.1097/JCMA.0000000000000270.

Xue, L., Jing, S., Miller, J.C., Sun, W., Li, H., Estrada-Franco, J.G., Hyman, J.M., Zhu, H., 2020. A data-driven network model for the emerging covid-19 epidemics in wuhan, toronto and italy. Mathematical Biosciences 326, 108391. URL: http://www.sciencedirect.com/science/article/pii/S0025556420300730, doi:https://doi.org/10.1016/j.mbs.2020.108391.

Yang, C., Sha, D., Liu, Q., Li, Y., Lan, H., Guan, W.W., Hu, T., Li, Z., Zhang, Z., Thompson, J.H., et al., 2020. Taking the pulse of covid-19: A spatiotemporal perspective. International journal of digital earth 13, 1186–1211.

CRediT authorship contribution statement

James Matuk: Conceptualization, Data curation, Methodology, Software, Visualization, Writing - Original draft preparation. **Xiaohan Guo:** Conceptualization, Methodology, Software, Writing - Original draft preparation.